(1 February 2008)

# ASYMPTOTIC BLOCKING PROBABILITIES IN LOSS NETWORKS WITH SUBEXPONENTIAL DEMANDS

YINGDONG LU AND ANA RADOVANOVIĆ,* *IBM T.J. Watson Research Center*

**Abstract**

The analysis of stochastic loss networks has long been of interest in computer and communications networks and is becoming important in the areas of service and information systems. In traditional settings, computing the well known Erlang formula for blocking probability in these systems becomes intractable for larger resource capacities. Using compound point processes to capture stochastic variability in the request process, we generalize existing models in this framework and derive simple asymptotic expressions for blocking probabilities. In addition, we extend our model to incorporate reserving resources in advance. Although asymptotic, our experiments show an excellent match between derived formulas and simulation results even for relatively small resource capacities and relatively large values of blocking probabilities.

**Keywords:** loss networks; subexponential distributions.
2000 Mathematics Subject Classification: Primary 60K25
Secondary 60J05;60K05;60K10

* Postal address: Mathematical Sciences Department, IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, {yingdong, aradovan}@us.ibm.com.

## 1. Introduction

The problem of satisfying a stream of customer (user) requirements from resources of finite capacities for some random processing time has long been present in many areas such as telephone and communication networks, inventory control (rental industry) and, recently, workforce management. For all of these applications, system dynamics can be described as follows. Requests for resources arrive according to some point process in time. If there are enough available (non-engaged) resources to satisfy their requirements at the moment of arrival, required resources are committed for some random time that represents their processing duration (holding time) after which they are released and become available to accommodate future requests. In the case of insufficient amount of available resources at the moment of its arrival, a request is lost. The previously described system is usually referred to as a *loss network*, and one of the commonly analyzed performance metrics is the blocking probability, i.e., probability that an incoming request is lost due to insufficient amount of available resources to satisfy its requirements.

Loss networks with fixed resource requirements have been intensively analyzed in the context of circuit-switched networks. Let requests require resources of $K < \infty$ different types for some random generally distributed processing time with finite mean. Furthermore, assume that requests belong to $M$ different classes characterized by their resource requirements, processing durations, arrival rates. Then, assuming that requests of different types arrive according to mutually independent Poisson processes, by PASTA property ([24]), blocking probability $B_l$ of an incoming request of type $1 \leq l \leq M$ is equal to the sum of probabilities of blocking states for $l$ type request and is computed using the generalized Erlang formula (e.g., see [15]),

i.e.,

$$B_l = 1 - G(\mathbf{C})^{-1} G(\mathbf{C} - \mathbf{A}\mathbf{e}_l),$$

where

$$G(\mathbf{C}) = \left( \sum_{\mathbf{n} \in \mathcal{S}(\mathbf{C})} \prod_{l=1}^{M} \frac{\rho_l^{n_l}}{n_l!} \right)$$

and

$$\mathcal{S}(\mathbf{C}) := \{\mathbf{n} \in \mathbb{Z}_+^M : \mathbf{A}\mathbf{n} \leq \mathbf{C}\}, \tag{1}$$

where $\mathbf{n} = (n_1, \ldots, n_M)$ and $\mathbf{C} = (C_1, \ldots, C_K)$. In the previous expressions $C_k$, $1 \leq k \leq K$, is capacity of resource type $k$, $\mathbf{A} = [A_{kl}]$ is a $K \times M$ matrix, where $A_{kl}$ represents the amount of resources of type $1 \leq k \leq K$ required by a request of type $1 \leq l \leq M$, and $\rho_l$, $1 \leq l \leq M$, represent traffic intensities of $l$ type requests (computed as $\rho_l = \lambda_l/\mu_l$, where $\lambda_l$ is the arrival rate of $l$ type requests and $1/\mu_l$ is the corresponding mean processing time). Furthermore, $e_l$ is a $M$ dimensional vector with the $l$th component equal to one and the rest equal to zero. In the case of a single resource type and a single request class with exponentially distributed processing times, blocking probability was first expressed by Erlang in 1917 (see [7]). Later on, it was shown that the Erlang formula holds under more general assumptions on call holding time distributions (see [20]) and in the case of Poisson arrivals with retrials (see [4]). It is noteworthy to point out the difference between the Erlang loss network and a queue with finite buffer. The two systems follow very different dynamics resulting in a different behavior and, therefore, their analysis (e.g., see [12] and [2]).

It is easy to see that the cardinality of the state space $\mathcal{S}(\mathbf{C})$ in (1) increases exponentially in the norm of vector $\mathbf{C}$, i.e., $|\mathbf{C}| \equiv \sum_{i=1}^{K} |C_i|$. It is shown in [18] that the calculation of $G(\mathbf{C})$ is a $\sharp P$-complete problem, which belongs to a class of problems that are at least as hard as

$NP$-complete problems. To this end, many approximation techniques for evaluating blocking probabilities in large loss networks have been proposed. One of the most popular ones is known as Erlang fixed point method. The main idea of this approximation is to assume that deficiencies of different resource types happen independently. The application of the Erlang fixed point method can be traced back as early as 50's (e.g., see [23]). In [14], Kelly studied the performance of the Erlang fixed point method and established its relation to a nonlinear optimization problem. He also proved uniqueness of the fixed point and its asymptotic exactness when resource capacities and arrival rates grow with the same rate (see [15]). Some of the related practical aspects of Kelly's analysis were investigated in [22]. The Erlang fixed point method is further refined in [25]. There are also many other types of approximations such as recursive algorithm in [13], or unified approach based on large deviations for all (light, critical and heavy) traffic regimes in [8]. Overall, except from the bounds in [8], these methods make use of the structural properties of the Erlang formula and, hence, largely rely on the Poisson assumption for call arrivals. Another restriction of the above models is that the amount of resource requirements are assumed to be fixed; in fact, it is assumed that they are $(0, 1)$ parameters in most of the cases considered. Meanwhile, we see in many applications that resource requirements could be highly variable and their distributions possibly long-tailed; for specific examples, see [10], [11] and [16]. Furthermore, more recently, loss networks models have been applied in the context of workforce management applications (see [19]), where requests behavior is even more volatile and extreme.

In this paper, we analyze loss networks that have renewal arrivals and random resource requirements. In particular, we assume that request arrivals follow a compound renewal process, with the corresponding holding times being arbitrarily distributed with finite mean, independent of each other and arrival points. In order to cope with variability in resource

requirements, we model them as subexponential random variables. We obtain a simple and explicit asymptotic expressions for blocking probabilities when capacities of resources grow. For the case of a single resource loss network, we show that the stationary blocking probability is approximately equal to the tail of the resource requirement distribution. In addition, we extend our results to allow advance reservations of resources. Finally, we investigate general (multiple resources and arbitrary topology) loss networks and show that the asymptotic blocking probability behaves as the tail of the heaviest-tailed resource requirement. Although asymptotic, our numerical experiments show an excellent accuracy of the derived formulas even for relatively small capacities and relatively large values of blocking probabilities, suggesting wide applicability of the obtained results.

Our paper is organized as follows. In Section 2, we introduce our model in the context of a single resource type. Then, in Subsection 2.1, we state and prove our main result in Theorem 1, while in Subsection 2.2, we extend it to the case of advance reservations. Further extension to the analysis of the stationary blocking probability in the case of general loss networks is stated and proved in Theorem 2 of Section 3. Our simulation experiments for some specific cases of arrival processes and resource requirements are presented in Section 4. Finally, we conclude our paper in Section 5. A discussion and the proof of existence of the stationary blocking probability is presented in the Appendix.

## 2. Systems with one resource type

Let requests for resources from a common resource pool of capacity $C < \infty$ arrive at time points $\{\tau_n, -\infty < n < \infty\}$ that represent a renewal process with rate $0 < \lambda < \infty$, i.e., $\mathbb{E}[\tau_n - \tau_{n-1}] = 1/\lambda$. At each point $\tau_n$, $B_n$ amount of resources is requested. If available

capacity is less than $B_n$, this request is rejected (blocked); otherwise, it is accepted and $B_n$ amount of resources will be occupied for the length of time $\theta_n$. Sequences $\{B_n\}$ and $\{\theta_n\}$ of i.i.d. random variables (r.v.) are assumed to be mutually independent and independent of the arrival points $\{\tau_n\}$; furthermore $\mathbb{E}\theta_n < \infty$ for all $n$. Let $B$ and $\theta$ denote random variables that represent $\{B_n\}$, $\{\theta_n\}$, i.e., $\mathbb{P}[B > x] = \mathbb{P}[B_n > x], \mathbb{P}[\theta > y] = \mathbb{P}[\theta_n > y]$, for any $n \in \mathbb{Z}$, $x \geq 0$ and $y \geq 0$.

In this paper, we assume that $B$ is a subexponential random variable, defined as follows (e.g., see [9]):

**Definition 1.** Let $\{X_i\}$ be a sequence of positive i.i.d. random variables with distribution function $F$ such that $F(x) < 1$ for all $x > 0$. Denote by $\bar{F}(x) = 1 - F(x)$, $x \geq 1$, the tail of $F$ and by $\bar{F}^{n*} = 1 - F^{n*}(x) = \mathbb{P}[X_1 + \cdots + X_n > x]$ the tail of the n-fold convolution of $F$. $F$ is subexponential distribution function, denoted as $F \in \mathcal{S}$, if one of the following equivalent conditions holds:

- $\lim_{x\to\infty} \frac{\bar{F}^{n*}(x)}{\bar{F}(x)} = n$ for some (all) $n \geq 2$,
- $\lim_{x\to\infty} \frac{\mathbb{P}[X_1+\cdots+X_n>x]}{\mathbb{P}[\max(X_1,\ldots,X_n)>x]} = 1$ for some (all) $n \geq 2$.

For a brief introduction to subexponential distributions the reader is referred to a recent survey [9]. This class of distributions is fairly large and well known examples include regularly varying (in particular Pareto), some Weibull, log-normal and "almost" exponential distributions.

Next, let $\mathcal{N}_n^{(C)}$ be the set of indices $i < n$ of resource requirements that arrive prior to $\tau_n$, are accepted, and are *still active* by time $\tau_n$. Furthermore, let $N_n^{(C)} \triangleq |\mathcal{N}_n^{(C)}|$ be a cardinality of set $\mathcal{N}_n^{(C)}$. Thus, the total amount of resources $Q_n^{(C)}$ that an arrival at time $\tau_n$ finds engaged can be expressed as $Q_n^{(C)} = \sum_{i\in\mathcal{N}_n^{(C)}} B_i$.

Our goal in this paper is to estimate the stationary blocking probability, i.e.,

$$\mathbb{P}[Q_n^{(C)} + B_n > C], \tag{2}$$

for large $C$. It can be shown that for the model introduced above there exists a unique stationary distribution for $Q_n^{(C)}$ and, therefore, the quantity in (2) is well defined. The proof of this result is based on constructing a Markov chain with general state space, of which $Q_n^{(C)}$ is a functional. Then, by using a discrete version of Theorem 1 from [20], we show that there exists a unique stationary distribution for the constructed Markov chain (and, therefore, $Q_n^{(C)}$) which is ergodic. Since this proof is not the main focus of this paper, we present it in the Appendix.

In this paper we use the following standard notation. For any two real functions $a(t)$ and $b(t)$ and fixed $t_0 \in \mathbb{R} \cup \{\infty\}$, let $a(t) \sim b(t)$ as $t \to t_0$ denote $\lim_{t\to t_0}[a(t)/b(t)] = 1$.

### 2.1. Blocking probability in a system with one resource type

In this section we estimate the stationary blocking probability $\mathbb{P}[Q_n^{(C)} + B_n > C]$ in a loss network with a single resource pool when its capacity $C$ grows large.

**Theorem 1.** *Let $\{B_n, -\infty < n < \infty\}$ be a sequence of subexponential random variables with finite mean. Then, the stationary blocking probability satisfies*

$$\mathbb{P}[Q_n^{(C)} + B_n > C] \sim \mathbb{P}[B > C] \ \ \textit{as} \ \ C \to \infty. \tag{3}$$

**Proof:** First, observe that a request will be lost if it requires more than the total capacity $C$ and, therefore,

$$\mathbb{P}[Q_n^{(C)} + B_n > C] \geq \mathbb{P}[B > C] \quad \text{for all } C > 0. \tag{4}$$

In order to prove the asymptotic upper bound for $\mathbb{P}[Q_n^{(C)} + B_n > C]$, we start by conditioning on the size of $B_n$ as

$$\mathbb{P}[Q_n^{(C)} + B_n > C] = \mathbb{P}[Q_n^{(C)} + B_n > C, B_n > C] + \mathbb{P}[Q_n^{(C)} + B_n > C, B_n \leq C]$$
$$\triangleq I_1 + I_2. \tag{5}$$

Note that $I_1$ is upper bounded by $\mathbb{P}[B > C]$. Next, we prove that $I_2 = o(\mathbb{P}[B > C])$ as $C \to \infty$. In view of the definition of $\mathcal{N}_n^{(C)}$ from above,

$$I_2 = \mathbb{P}\left[\sum_{i \in \mathcal{N}_n^{(C)}} B_i + B_n > C, B_n \leq C\right]. \tag{6}$$

Observe that for $i \in \mathcal{N}_n^{(C)}$, $B_i$s are mutually dependent which makes direct analysis of the expression in (6) complex. For that reason, we sample the original process of arrivals at points $\tau_i$ at which the requested amount of resources $B_i$ is smaller or equal to $C$ and observe another system of unlimited capacity with the sampled arrivals. Let $\mathcal{N}_{n,s}$ be a set of request indices $i < n$ that belong to the sampled process and are still active at time $\tau_n$, i.e.,

$$\mathcal{N}_{s,n} = \{i < n | B_i \leq C, \theta_i > \tau_n - \tau_i\}.$$

Note that the sampled process is renewal as well with rate $\lambda\mathbb{P}[B \leq C]/\mathbb{P}[B > C]$ and that resource requirements $B_i$, $i \in \mathcal{N}_{s,n}$, are mutually independent. Furthermore, since $\mathcal{N}_n^{(C)} \subset \mathcal{N}_{s,n}$, we can upper bound $I_2$ in (6) by the probability that the total amount of required resources in a new system exceeds capacity $C$, i.e.,

$$I_2 \leq \mathbb{P}\left[\sum_{i \in \mathcal{N}_{s,n}} B_i + B_n > C, B_n \leq C\right]. \tag{7}$$

Now, in view of the results derived in [6] for every integer $n$ and i.i.d. subexponential random variables $B_1, \ldots, B_n$, $\mathbb{P}[\sum_{i=1}^n B_i > C] \sim \mathbb{P}[\max(B_1, B_2, \ldots, B_n) > C]$ as $C \to \infty$,

implying asymptotic relation

$$\mathbb{P}\left[\sum_{i=1}^{n} B_i > C, B_i \leq C \text{ for every } 1 \leq i \leq n\right] = o(\mathbb{P}[B > C]) \text{ as } C \to \infty.$$

In order to show that $n$ can be replaced by $N_{s,n}$ in the above inequality, we need to integrate it with respect to the density of $N_{s,n}$, i.e.,

$$\mathbb{P}\left[\sum_{i\in\mathcal{N}_{s,n}\cup\{n\}} B_i > C, B_i \leq C \text{ for every } i \in \mathcal{N}_{s,n}\cup\{n\}\right]$$
$$= \sum_{k=0}^{\infty} \mathbb{P}[N_{s,n} = k]\mathbb{P}\left[\sum_{i=1}^{k+1} B_i > C, B_i \leq C \text{ for every } i = 1, \ldots, k+1\right].$$

Note that on the left hand side of the previous equation index $i$ can take negative values. Next, due to the lemma stated by Kesten (see Lemma 7, pp.149 of [3]), for any $\epsilon > 0$ there exists a positive constant $K(\epsilon)$ such that

$$\frac{\mathbb{P}[\sum_{i=1}^{k} B_i > C, B_i \leq C \text{ for every } 1 \leq i \leq k]}{\mathbb{P}[B > C]} \leq \frac{\mathbb{P}[\sum_{i=1}^{k} B_i > C]}{\mathbb{P}[B > C]} \leq K(\epsilon)(1+\epsilon)^k,$$

for any integer $k$ and all capacity values $C < \infty$. Then, since the probability generating function $\mathbb{E}z^{N_{s,n}}$ is finite for any $z \in \mathbb{C}$ (see Theorem 1 in [21] and Theorem 5 in [17] for the detailed proof), we have $\sum_{k=0}^{\infty} \mathbb{P}[N_{s,n} = k](1+\epsilon)^k < \infty$. Therefore, by applying the dominated convergence theorem, we conclude that

$$\lim_{C\to\infty} \frac{\mathbb{P}\left[\sum_{i\in N_{s,n}} B_i + B_n > C, B_i \leq C \text{ for every } i \in \mathcal{N}_{s,n}\cup\{n\}\right]}{\mathbb{P}[B > C]}$$
$$= \lim_{C\to\infty} \sum_{k=0}^{\infty} \frac{\mathbb{P}[N_{s,n} = k]\mathbb{P}\left[\sum_{i=1}^{k+1} B_i > C, B_i \leq C \text{ for every } 1 \leq i \leq k+1\right]}{\mathbb{P}[B > C]}$$
$$= 0, \tag{8}$$

which in conjunction with (5) and (4), completes the proof of this theorem. ◇

**Remark:** It may appear surprising that the performance of the loss network from above does not depend on engagement durations, as long as they have finite mean. In addition, the result is quite general and provides the asymptotic result for a large (subexponential) class of possible resource requirement distributions.

### 2.2. Advance reservations

Using the result of Theorem 1 and observations from the previous remark, we extend the loss networks model to allow requests to become effective with some delay with respect to the moments of their arrivals. In particular, a request that arrives at time $\tau_n$ and requires $B_n$ amount of resources for some random time $\theta_n$ starting from the moment $\tau_n + D_n$ is accepted if previously admitted resource requirements allow that; otherwise, it is rejected. In other words, a request arriving at $\tau_n$ is lost if at any moment of time in interval $(\tau_n + D_n, \tau_n + D_n + \theta_n)$ the total amount of active requirements requested prior to $\tau_n$ exceeds $C - B_n$. First, note that $B_n > C$ implies the loss of $n$th request and, therefore, it is straightforward to conclude that the blocking probability in the system with advance reservations can be lower bounded by $\mathbb{P}[B > C]$.

Next, we discuss the idea behind proving the upper bound on the blocking probabilities. By applying sample path arguments one can show that, at any moment of time, the amount of active resources in the previously described system with advance reservations can be bounded from above by the amount of active resources in another system of unlimited capacity, without advance reservations, with resource holding times $D_n + \theta_n$ for every $n$, and with requests for resources being sampled from the original process $\{B_n\}$ whenever the corresponding requirements are less or equal to $C$. Equivalently, the blocking probability in the system with

advance reservations can be bounded from above by

$$\mathbb{P}\left[\sum_{i\in\mathcal{N}_{s,n}^{(C)}(\theta+D)} B_i + B_n > C\right],$$

where $\mathcal{N}_{s,n}^{(C)}(\theta+D)$ is a set of request indices $i < n$ that are active at time $\tau_n$, whose requirements are less or equal to $C$ and holding times last throughout the interval $(\tau_i, \tau_i + D_i + \theta_i)$, assuming that there is an unlimited resource capacity.

Finally, by using the previous discussion, the properties of $\{B_n\}$, $\{\theta_n\}$ and $\{\tau_n\}$ as introduced at the beginning of this section, assuming that reservation times $\{D_n\}$, $\mathbb{E}D_n < \infty$, are i.i.d. and independent from $\{B_n\}$, $\{\theta_n\}$ and $\{\tau_n\}$, and applying the identical arguments as in the proof of Theorem 1, we obtain the following result:

**Corollary 1.** *The blocking probability in the system with advance reservations approaches* $\mathbb{P}[B > C]$ *as* $C \to \infty$.

## 3. Acquiring resources of different types (loss networks case)

Assume that there are $K \in \mathbb{N}$ resource types with capacities $C_1, \ldots, C_K$. Again, requests arrive at $\{\tau_n, -\infty < n < \infty\}$, which represent a renewal process with rate $0 < \lambda = 1/\mathbb{E}[\tau_1 - \tau_0] < \infty$. There are $M < \infty$ request types and, given an arrival, the request is of type $l$, $1 \le l \le M$, with probability $p_l$, $p_1 + \cdots + p_M = 1$, independent from $\{\tau_n\}$. We will use random variables $J_n \in \{1, 2, \ldots, M\}$ to denote the type of the request arriving at $\tau_n$. Furthermore, let $B_n^{(J_n,1)}, \ldots, B_n^{(J_n,K)}$ represent amounts of required resources of each type at time $\tau_n$ and let $\theta_n^{(J_n)}$, $\mathbb{E}\theta_n^{(J_n)} < \infty$, be the corresponding random duration. We assume that sequences $\{(B_n^{(J_n,1)}, \ldots, B_n^{(J_n,K)})\}, \{\theta_n^{(J_n)}\}$ are mutually independent and independent from $\{\tau_n\}$. Given the event $\{J_n = l\}$, resource requirements $B_n^{(l,i)}$, $1 \le i \le K$, are mutually

independent nonnegative random variables drawn from distributions $F_{l,i}$, $1 \leq i \leq K$; if a request does not require resources of type $i$ then $B_n^{(l,i)} = 0$ a.s., $-\infty < n < \infty$. Only if there is enough capacity available, the request arriving at time $\tau_n$ will be accepted and all of the engaged resources will be occupied for the duration of $\theta_n^{(J_n)}$; otherwise, the request is rejected.

Our goal is to estimate the blocking probability in a system described above. Define $Q_n^{(1)}, \ldots, Q_n^{(K)}$ to be amounts of resources of each type that a request arriving at time $\tau_n$ finds engaged. Note that $Q_n^{(i)}$, $1 \leq i \leq K$, are mutually dependent and, as pointed out in the Introduction, it is hard to compute the blocking probability of this system explicitly. Using analogous arguments as in the case of a single resource type (see the Appendix), one can show that the stationary distribution of $Q_n^{(i)}$, $1 \leq i \leq K$, exists. Probability that the request arriving at time $\tau_n$ is blocked equals to

$$\mathbb{P}[\cup_{1 \leq i \leq K} \{Q_n^{(i)} + B_n^{(J_n,i)} > C_i\}], \tag{9}$$

and our goal again is to estimate its value as $\min_i C_i$ grows large.

Asymptotic estimates derived in this section hold under the following assumption:

**Assumptions:** For each resource type $1 \leq i \leq K$, let $\mathcal{L}_i$ and $\mathcal{H}_i$ be two disjoint sets of request types ($|\mathcal{L}_i \cup \mathcal{H}_i| = M$) satisfying:

- Assume that there exists at least one resource type that is accessed by subexponentially distributed resource requirements, which implies $|\mathcal{H}_i| > 0$ for some $1 \leq i \leq K$;
- For every $l \in \mathcal{H}_i \neq \emptyset$, there exists a subexponential distribution $F_i \in \mathcal{S}$ such that $\bar{F}_{l,i}(x) \sim c_{l,i}\bar{F}_i(x)$ as $x \to \infty$ with $c_{l,i} > 0$;

- There exists a subexponential random variable $L \in \mathcal{S}$ that satisfies

$$\mathbb{P}[L > x] \geq \max_{1 \leq i \leq K, l \in \mathcal{L}_i} \mathbb{P}[B_n^{(J_n,i)} > x | J_n = l] \text{ for all } x > 0,$$

and $\mathbb{P}[L > x] = o(\bar{F}_i(x))$ as $x \to \infty$ for all $i \in \{j | \mathcal{H}_j \neq \emptyset\}$.

**Remark:** In the preceding assumptions, we require the resource requirement distributions to be asymptotically comparable. For each $1 \leq i \leq K$, $\mathcal{H}_i$ contains tail dominant subexponential distributions that are asymptotically proportional to each other. On the other hand, the only assumption imposed on the distributions in $\mathcal{L}_i$, $1 \leq i \leq K$, is that there is a subexponential tail that asymptotically dominates them. This asymptotic tail comparability is necessary for our main result to hold. In particular, these conditions are extensively used in (16) - (21) of the proof of Theorem 2.

Next, we prove the following lemma that investigates summations of random variables with different tail distributions.

**Lemma 1.** *Suppose that $X_i, 1 \leq i \leq n$, are independent random variables with corresponding tail distributions $\bar{F}_i(x)$, $1 \leq i \leq n$. If there exists $F \in \mathcal{S}$ such that $\bar{F}_i(x) \sim c_i \bar{F}(x)$ as $x \to \infty$ with $c_i \geq 0, 1 \leq i \leq n$, and $\sum_{i=1}^n c_i > 0$, then the following asymptotic relation holds:*

$$\mathbb{P}\left[\sum_{i=1}^n X_i > x, X_i \leq x, 1 \leq i \leq n\right] = o(\bar{F}(x)) \text{ as } x \to \infty. \tag{10}$$

**Proof:** Note that

$$\mathbb{P}\left[\sum_{i=1}^n X_i > x\right] = \mathbb{P}\left[\sum_{i=1}^n X_i > x, X_i \leq x, 1 \leq i \leq n\right] + \mathbb{P}\left[\sum_{i=1}^n X_i > x, \cup_{i=1}^n \{X_i > x\}\right].$$

Then, the previous expression, $\cup_{i=1}^n \{X_i > x\} \subset \{\sum_{i=1}^n X_i > x\}$, independence of $X_i$s, as well as Lemmas 4.2 and 4.5 of [1], imply (10). $\diamond$

First, we estimate the asymptotic lower bound for the expression in (9). By using our model assumptions, $\{B_n^{(J_n,i)} > C_i\} \subset \{Q_n^{(i)} + B_n^{(J_n,i)} > C_i\}$ and independence, we obtain

$$\mathbb{P}[\cup_{1\leq i\leq K}\{Q_n^{(i)} + B_n^{(J_n,i)} > C_i\}] \geq \mathbb{P}[\cup_{1\leq i\leq K}\{B_n^{(J_n,i)} > C_i\}] \sim \sum_{i=1}^{K}\sum_{l\in\mathcal{H}_i} p_l \bar{F}_{l,i}(C_i), \tag{11}$$

as $\min_i C_i \to \infty$.

Next, we estimate the asymptotic upper bound for the expression in (9). Using the union bound yields

$$\mathbb{P}[\cup_{1\leq i\leq K}\{Q_n^{(i)} + B_n^{(J_n,i)} > C_i\}] \leq \sum_{i=1}^{K}\mathbb{P}[Q_n^{(i)} + B_n^{(J_n,i)} > C_i]. \tag{12}$$

Similarly as in (7) of Theorem 1, for each resource $1 \leq i \leq K$,

$$\mathbb{P}[Q_n^{(i)} + B_n^{(J_n,i)} > C_i] \leq \mathbb{P}\left[\sum_{l\in\mathcal{L}_i}\sum_{j\in\mathcal{N}_{s,n}^{(l,C_i)}} B_j^{(l,i)} + \sum_{l\in\mathcal{H}_i}\sum_{j\in\mathcal{N}_{s,n}^{(l,C_i)}} B_j^{(l,i)} + B_n^{(J_n,i)} > C_i\right], \tag{13}$$

where $\mathcal{N}_{s,n}^{(l,C_i)}$, $1 \leq l \leq M$, are sets of indices $j < n$ defined as

$$\mathcal{N}_{s,n}^{(l,C_i)} \triangleq \{j < n | J_j = l, B_j^{(l,i)} \leq C_i, \theta_j^{(l)} > \tau_n - \tau_j\}.$$

In the previous expressions we bounded the amount of allocated resources that are active at time $\tau_n$ by the corresponding quantity in another system of infinite capacity where the corresponding request process is sampled from the original $\{B_n^{(J_n,i)}\}$, $1 \leq i \leq K$, whenever the corresponding requirements are less than or equal to $C_i$, $1 \leq i \leq K$.

In the rest of the proof, we derive an asymptotic estimate for the expression in (13). After conditioning on $\{N_{s,n}^{(1,C_i)} = n_1, \ldots, N_{s,n}^{(M,C_i)} = n_M\}$ ($N_{s,n}^{(l,C_i)} \triangleq |\mathcal{N}_{s,n}^{(l,C_i)}|$, $1 \leq l \leq M$), we

obtain

$$\mathbb{P}[Q_n^{(i)} + B_n^{(J_n,i)} > C_i]$$

$$\leq \sum_{0\leq n_1,\ldots,n_M<\infty} \mathbb{P}[N_{s,n}^{(1,C_i)} = n_1, \ldots, N_{s,n}^{(M,C_i)} = n_M]$$

$$\times\, \mathbb{P}\left[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n_l} B_{(j)}^{(l,i)} + \sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n_l} B_{(j)}^{(l,i)} + B_n^{(J_n,i)} > C_i,\ B_{(j)}^{(l,i)} \leq C_i, 1\leq j\leq n_l, 1\leq l\leq M\right], \tag{14}$$

where $B_{(j)}^{(l,i)} \stackrel{d}{=} B_k^{(l,i)}$, $k \in \mathcal{N}_{s,n}^{(l,C_i)}$, $j = 1,\ldots,n_l$, are independent replicas of requests in $\mathcal{N}_{s,n}^{(l,C_i)}$. Next, after conditioning on $\{J_n = m\}$, $m = 1,\ldots,M$, and then on $B_n^{(m,i)}$ being smaller or larger than $C_i$, we can further upper bound the conditional blocking probability in (14) as

$$\mathbb{P}\left[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n_l} B_{(j)}^{(l,i)} + \sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n_l} B_{(j)}^{(l,i)} + B_n^{(J_n,i)} > C_i,\ B_{(j)}^{(l,i)} \leq C_i, 1\leq j\leq n_l, 1\leq l\leq M\right] \leq$$

$$\sum_{m=1}^{M} p_m \mathbb{P}\left[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n_l} B_{(j)}^{(l,i)} + \sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n_l} B_{(j)}^{(l,i)} + B_n^{(m,i)} > C_i,\ B_{(j)}^{(l,i)} \leq C_i, 1\leq j\leq n_l, 1\leq l\leq M, B_n^{(m,i)} \leq C_i\right]$$

$$+ \sum_{m=1}^{M} p_m \mathbb{P}[B_n^{(m,i)} > C_i]. \tag{15}$$

Thus, the probabilities in the first term on the right hand side of the previous expression can be expressed in the form

$$\mathbb{P}\left[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n_l'} B_{(j)}^{(l,i)} + \sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n_l'} B_{(j)}^{(l,i)} > C_i,\ B_{(j)}^{(l,i)} \leq C_i, 1\leq j\leq n_l', 1\leq l\leq M\right], \tag{16}$$

where $n_l' = n_l$ for $l \neq m$ and $n_l' = n_l + 1$ for $l = m$.

Next, in order to estimate the asymptotic upper bound of the term in (16), Assumptions enable us to distinguish between two cases: (i) $\mathcal{H}_i = \emptyset$ or $\sum_{l\in\mathcal{H}_i} n_l' = 0$, and (ii) $\mathcal{H}_i \neq \emptyset$ and $\sum_{l\in\mathcal{H}_i} n_l' > 0$.

(i): If $\mathcal{H}_i = \emptyset$ or $\sum_{l\in\mathcal{H}_i} n'_l = 0$, we have that probability in (15) can be upper bounded as

$$\mathbb{P}\left[\sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n'_l} B_{(j)}^{(l,i)} > C_i\right] \leq \mathbb{P}\left[\sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n'_l} L_{(j)}^{(l,i)} > C_i\right],$$

where in the inequality above we used Assumptions and introduced $L_{(j)}^{(l,i)}$ to be independent r.v.s equal in distribution to $L$. Hence, since $L_{(j)}^{(l,i)}$ are subexponential, we obtain

$$\lim_{C_i\to\infty} \frac{\mathbb{P}\left[\sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n'_l} B_{(j)}^{(l,i)} > C_i\right]}{\mathbb{P}[L > C_i]} \leq \sum_{l\in\mathcal{L}_i} n'_l. \tag{17}$$

(ii): If $\mathcal{H}_i \neq \emptyset$ and $\sum_{l\in\mathcal{H}_i} n'_l > 0$, using Assumptions and Lemma 1, we derive the following asymptotic upper bound

$$\mathbb{P}\left[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n'_l} B_{(j)}^{(l,i)} + \sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n'_l} B_{(j)}^{(l,i)} > C_i,\ B_{(j)}^{(l,i)} \leq C_i, 1\leq j\leq n'_l, 1\leq l\leq M\right] = o(\bar{F}_i(C_i)), \tag{18}$$

as $C_i \to \infty$.

Thus, in (16)-(18) we obtained upper bounds and their asymptotic estimates for the conditional blocking probabilities in the first term of (15) that hold for any finite nonnegative integers $n_1, \ldots, n_M$. Thus, in view of (14), in order to estimate an asymptotic upper bound of $\mathbb{P}[Q_n^{(i)} + B_n^{(J_n,i)} > C_i]$, we need to integrate probabilities in (16) with respect to densities of r.v.s $N_{s,n}^{(l,C_i)}$, $l = 1, \ldots, M$. In this regard, note that in the case where $\mathcal{H}_i \neq \emptyset$, by Assumptions, the term in (16) can be upper bounded as

$$\begin{aligned}&\mathbb{P}\left[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n'_l} B_{(j)}^{(l,i)} + \sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n'_l} B_{(j)}^{(l,i)} > C_i,\ B_{(j)}^{(l,i)} \leq C_i, 1\leq j\leq n'_l, 1\leq l\leq M\right]\\ &\leq \mathbb{P}\left[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n'_l} B_{(j)}^{(l,i)} + \sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n'_l} L_{(j)}^{(l,i)} > C_i\right],\end{aligned} \tag{19}$$

where, as before, $L_{(j)}^{(l,i)}$ are independent r.v.s equal in distribution to $L$. Furthermore, since $\mathbb{P}[L > x] = o(\bar{F}_i(x))$ as $x \to \infty$, there exists a large enough finite integer $H$ such that

$\mathbb{P}[L > x] \leq H\bar{F}_i(x)$ for all $x \geq 0$. Therefore, for any $x \geq 0$, one can write

$$\mathbb{P}[L > x] \leq H\bar{F}_i(x) = \mathbb{P}\left[\cup_{1\leq r\leq H}\{\hat{B}_r^{(i)} > x\}\right] \leq \mathbb{P}\left[\sum_{r=1}^{H}\hat{B}_r^{(i)} > x\right], \tag{20}$$

where $\hat{B}_r^{(i)}$, $1 \leq r \leq H$, are independent r.v.s having cumulative distribution function $F_i$. Now, in view of (20), each of random variables $L_{(j)}^{(l,i)}$ in (19) can be stochastically upper bounded by a random variable that is equal in distribution to $\sum_{r=1}^{H}\hat{B}_r^{(i)}$. Thus, if we introduce $Y_j$, $j \geq 1$, to be independent r.v.s equal in distribution to $\sum_{r=1}^{H}\hat{B}_r^{(i)}$, we obtain

$$\mathbb{P}\left[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n_l'}B_{(j)}^{(l,i)} + \sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n_l'}L_{(j)}^{(l,i)} > C_i\right] \leq \mathbb{P}\left[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n_l'}B_{(j)}^{(l,i)} + \sum_{j=1}^{\sum_{l\in\mathcal{L}_i}n_l'}Y_j > C_i\right],$$

which in conjunction with point (b) of Lemma 4.2 in [1] implies that for any $\epsilon > 0$ there exist a finite constant $K_\epsilon$ such that

$$\begin{aligned}\mathbb{P}\left[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n_l'}B_{(j)}^{(l,i)} + \sum_{l\in\mathcal{L}_i}\sum_{j=1}^{n_l'}L_{(j)}^{(l,i)} > C_i\right] &\leq \mathbb{P}\left[\sum_{l\in\mathcal{H}_i}\sum_{j=1}^{n_l'}B_{(j)}^{(l,i)} + \sum_{j=1}^{\sum_{l\in\mathcal{L}_i}n_l'}Y_j > C_i\right] \\ &\leq K_\epsilon(1+\epsilon)^{\sum_{l\in\mathcal{H}_i}n_l'+\sum_{l\in\mathcal{L}_i}n_l'}\bar{F}_i(C_i), \end{aligned}\tag{21}$$

for any $C_i < \infty$. Similarly, in cases where $\mathcal{H}_i = \emptyset$, we could apply the stochastic dominance $B_{(j)}^{(l,i)} \overset{d}{\leq} L_{(j)}^{(l,i)}$, $l \in \mathcal{L}_i$, where $L_{(j)}^{(l,i)}$ are, as before, independent subexponential random variables equal in distribution to $L$. Then, by Kesten's lemma (see Lemma 7 on page 149 of [3]), the analogous bound to the one in (21) follows.

Finally, since (21) bounds uniformly probabilities in (16) for all $C_i < \infty$ and $n_l'$, $1 \leq l \leq M$, in conjunction with (15), (14), $N_{s,n}^{(l,C_i)} \leq N_n^{(l,\infty)}$ a.s. and existence of $\mathbb{E}z^{N_n^{(l,\infty)}}$ for all $z \in \mathbb{C}$, $1 \leq l \leq M$, (see Theorem 1 in [21] and Theorem 5 in [17]), one can apply the dominated convergence theorem and conclude

$$\lim_{C_i\to\infty}\frac{\mathbb{P}[Q_n^{(i)} + B_n^{(J_n,i)} > C_i]}{\sum_{l\in\mathcal{H}_i}p_l\bar{F}_{l,i}(C_i)} \leq 1[\mathcal{H}_i \neq \emptyset].$$

Next, by adding asymptotic estimates for all $1 \leq i \leq K$, in conjunction with (11), we complete the proof of the following result:

**Theorem 2.** *For the request model introduced in this section, under the conditions imposed by Assumptions, the stationary blocking probability for general loss networks satisfies*

$$\mathbb{P}[\cup_{1\le i\le K}\{Q_n^{(i)}+B_n^{(J_n,i)}>C_i\}]\sim\sum_{i=1}^{K}\sum_{l\in\mathcal{H}_i}p_l c_{l,i}\bar{F}_i(C_i)\ \ \textit{as}\ \ \min_i C_i\to\infty.$$

## 4. Numerical examples

In this section, with two simulation experiments, we demonstrate the accuracy of our asymptotic formulas, proved in Theorems 1 and 2. Our goal is to show that even though our results are asymptotic, the derived estimates match experiments with high accuracy even for systems with finite support demand distributions and moderately large capacities.

In each experiment, in order for the system to reach stationarity, we let the first $10^8$ arrivals to be a warm-up time. By repeating many experiments, we observe that longer warm-up times do not lead to improved results. Then, we count the number of blocked requests among next $10^9$ arrivals. In both of the experiments below, measurements are conducted for capacities $C=500+100j, 0\le j\le 9$, where the starting value of $C=500$ is set to be slightly larger than the effective systems load $\lambda\mathbb{E}[\theta_n]\mathbb{E}[B_n]$. Simulation results are presented by symbol "o" in Figures 1 and 2, while our approximations, estimates obtained in Theorems 1 and 2, are the solid lines on the same figures. Note that in order to emphasize the difference and to observe a range of blocking probabilities we are trying to estimate, we present base 10 logarithm of the obtained values.

**Example 1** Consider the case of a single resource type of capacity $C$. Let requests for resources arrive at Poisson time points with rate $\lambda=1$. In addition, we assume that engagement durations are exponentially distributed with mean $1/\mu=1$. Next, let request requirements $B_n$ be drawn from a finite support distribution, where $\mathbb{P}[B_n=i]=\frac{0.3}{i^{1.5}}$, $1\le i\le 1999$, and

$\mathbb{P}[B_n = 2000] = 1 - \mathbb{P}[B_n < 2000]$ (power law distribution). Effective load in this example is $\lambda\mathbb{E}[\theta_n]\mathbb{E}[B_n] \approx 485.8$. Experimental results are presented in Figure 1. Even though we start measuring rejections at capacities that are slightly larger than the mean requirement value, our approximation $\mathbb{P}[B_n > C]$ is very close to experimental results. In particular, the relative approximation error is less than $1\%$ for $C = 500$, and for capacity values larger or equal to $C = 1400$ this error is less than $0.3\%$.

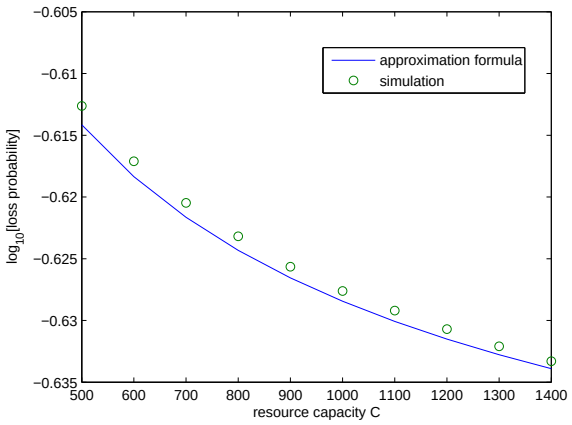


FIGURE 1: Illustration for Example 1

**Example 2** In this example, we consider the case of two resource and two request types. Furthermore, we assume that resource capacities are the same $C = C_1 = C_2$. The frequencies of requests of types 1 and 2 are $p_1 = 0.3$ and $p_2 = 0.7$ respectively. Assume that the arrival points are separated by a fixed, unit length of time, i.e., $\tau_n - \tau_{n-1} = 1$ for all $n$. Type 1 request durations satisfy $\theta_i^{(1)} \sim exp(4)$ and type 2 request holding times are drawn from the uniform distribution on $[0, 40]$, i.e., $\theta_i^{(2)} \sim Unif([0, 40])$. Resource requirements corresponding to engagements of type 1 are distributed as $\mathbb{P}[B^{1,1} = 1] = 0.8$, $\mathbb{P}[B^{1,1} = i] = 0.15e^{-\sqrt{i}}$, $2 \leq i \leq 1999$ and $\mathbb{P}[B^{1,1} = 2000] = 1 - \mathbb{P}[B^{1,1} < 2000]$ for the type 1 resources, and

$\mathbb{P}[B^{1,2} = 50] = 1$ for type 2 resources. Requests of type 2 require resources according to $\mathbb{P}[B^{2,1} = i] = geom^{i-1}(1 - geom)$, $1 \leq i \leq 1999$, $\mathbb{P}[B^{2,1} = 2000] = 1 - \mathbb{P}[B^{2,1} < 2000]$, where $geom = 0.6$ for resources of type 1, and $\mathbb{P}[B^{2,2} = i] = \frac{0.3}{i^{1.5}}$, $1 \leq i \leq 1999$, $\mathbb{P}[B^{2,2} = 2000] = 1 - \mathbb{P}[B^{2,2} < 2000]$ for type 2 resources. Our asymptotic results suggest that the blocking probability should be characterized by the heaviest tailed demand distributions. The results of this experiment are presented in Figure 2. As in the previous case, we obtain a very accurate agreement between our approximation and the simulation. The relative approximation error in this case does not exceed $2\%$ and is getting smaller as resource capacities grow.

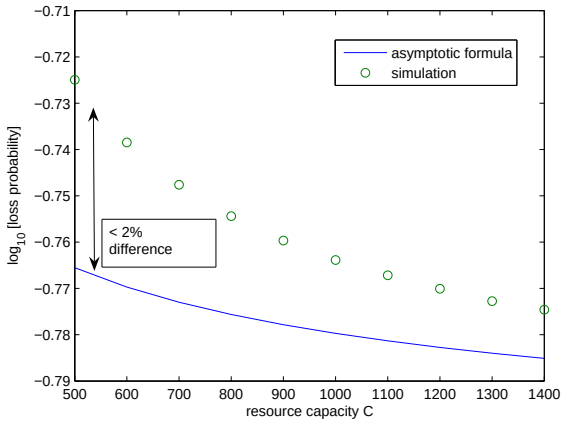


FIGURE 2: Illustration for Example 2

**Remark:** (i) We would like to point out that the accuracy of experimental results directly depends on the approximation errors (7) and (18), depending on the simulated scenarios. These errors highly depend on the tail properties of the resource requirements distributions. More specifically, under fairly general assumptions, the heavier the dominant tail of the resource requirement distribution is, the smaller would be the relative approximation error. For detailed explanations, a reader is referred to Section 1.3.2 of [5]. (ii) Note that our main results

estimate the stationary blocking probability and, as we commented earlier, are indifferent to distributional properties of holding times. For that reason, as long as one can claim that the measurements are conducted in stationarity, the transience should not affect experimental results.

## 5. Concluding remarks

In this paper, we consider loss networks with reusable resources and finite resource capacities and estimate the probability that a request is rejected due to insufficient amount of resources at points of their arrivals. Assuming a renewal process of request arrivals, subexponential resource requirements and generally distributed activity durations, we show that the asymptotic blocking probability for a wide class of analyzed systems can be fully estimated using resource requirement distribution, independent from other system's properties. In particular, we show that the blocking probability behaves as the asymptotically dominant tail of the resource requirement distribution.

The model we study can be applied to a wide range of applications. Historically, loss networks (in particular, Erlang loss networks) are widely used for modeling communication networks. Later, through the development of new services applications such as workforce management with similar modeling properties, the importance of accurately estimating blocking probabilities of general loss networks has become significant. In this regard, we investigate loss networks with various request types and possibly highly variable random amounts of required resources. In addition, we research the possibility of incorporating random advance reservations for incoming requests. These results should be of great interest to an emerging research community. Although our results are intended mainly for qualitative purposes, nu-

merical examples demonstrate an excellent match between derived formulas and simulated systems performance, hence strongly suggesting their application.

## Appendix

In this section we prove the existence of the stationary blocking probabilities in (2). Using the model description from Section 2, we observe the system at the moments of request arrivals. Then, we define a discrete time process $X_n \triangleq (N_n^{(C)}, B_i, E_i, i = 1, 2, \cdots, N_{0,n}^{(C)})$, where $E_i$s represent times that elapsed in processing requests in the system by time $\tau_n$; furthermore, $N_{0,n}^{(C)}$ is the number of active requests at the moment of $n$th arrival. Note that $\{X_n\}_{n\geq 0}$ is a discrete time Markov chain with state space $\Omega \triangleq \mathbb{N}_0 \times \ell_\infty \times \ell_\infty$, where $\ell_\infty$ denotes the Banach space of the infinite sequence of real numbers equipped with the supreme norm; let $\omega_0 \in \Omega$ denote the state with no active requests. We start observing the system at the moment $\tau_0$ of 0th arrival and denote the initial state by $X_0 = (N^0, B_i^0, E_i^0, i = 1, \ldots, N^0)$ drawn from some arbitrary distribution $P_0$, where $\mathbb{E}B_i^0 < \infty$, $\mathbb{E}\theta_i^0 < \infty$. Next, define $\mathcal{F}$ to be the Borel field of $\Omega$, and let $P_n(x_0, A)$, $x_0 \in \Omega$, $A \in \mathcal{F}$, represent a transition probability of the Markov chain $X_n$ into set $A$ in time $n$, starting from state $x_0$. Let $P_n$ be the probability distribution of $X_n$.

Now, in order to prove the existence of a unique stationary distribution for the Markov chain $\{X_n\}$, we use a discrete version of Theorem 1 in [20], which we state next for reasons of completeness.

**Theorem:** A Markov chain homogeneous in time has a unique stationary distribution which is ergodic if, for any $\epsilon > 0$, there exists a measurable set $S$, a probability distribution $R$ in $\Omega$, and $n_1 >$, $k > 0$, $K > 0$ such that

- $kR(A) \leq P_{n_1}(x, A)$ for all points $x \in S$ and measurable sets $A \subset S$; for any initial distribution $P_0$ there exists $n_0$ such that for any $n \geq n_0$
- $P_n(S) \geq 1 - \epsilon$,
- $P_n(A) \leq KR(A) + \epsilon$ for all measurable sets $A \subset S$.

**Proof:** The proof follows identical arguments as in [20] translated into discrete setting.

Next, we need to show that Theorem 1 holds for the process investigated in this paper; in particular, we will consider a common resource pool case. The proof follows the similar reasoning as in Theorems 4 and 5 of [20].

Define set $S(\psi, \beta, \delta)$ as

$$S(\psi, \beta, \delta) \triangleq \{N_{0,n}^{(C)} \leq \psi, 0 \leq B_i \leq \beta, 0 \leq E_i \leq \delta, i \in \mathcal{N}_{0,n}^{(C)}\}$$

for some positive finite constants $\psi, \beta, \delta$.

Now, we show that for any $\epsilon > 0$, there exists $S(\psi, \beta, \delta) \subset \Omega$ such that for any initial distribution $P_0$ there exists $n_0$ such that for all $n \geq n_0$

$$P_n(S(\psi, \beta, \delta)) \geq 1 - \epsilon. \tag{22}$$

Note that

$$P_n(\bar{S}(\psi, \beta, \delta)) \leq \mathbb{P}[\cup_{i \in \mathcal{N}_{0,n}^{(C)}} \{\theta_i > \delta\}, N_{0,n}^{(C)} \leq \psi] + \mathbb{P}[\cup_{i \in \mathcal{N}_{0,n}^{(C)}} \{B_i > \beta\}, N_{0,n}^{(C)} \leq \psi] + \mathbb{P}[N_{0,n}^{(C)} > \psi]$$

$$\leq \psi \mathbb{P}[\theta_i > \delta] + \psi \mathbb{P}[B_i > \beta] + \mathbb{P}[N_{a,n}^{(C)} + N_{0,n}^{0} > \psi], \tag{23}$$

where $N_{a,n}^{(C)}$ represents the number of active requests at $\tau_n$ that originated from $n$ arrivals at $\tau_0, \ldots \tau_{n-1}$, and the rest of active requests at $\tau_n$, $N_{0,n}^{0} = N_{0,n}^{(C)} - N_{a,n}^{(C)}$ are those that were

active at the initial point $\tau_0$ and are still processed at the moment of $n$th arrival. Next, since

$$\begin{aligned}
&\mathbb{P}[N_{a,n}^{(C)} + N_{0,n}^0 > \psi] \leq \mathbb{P}\left[N_{a,n}^{(C)} > \frac{\psi}{2}\right] + \mathbb{P}\left[N_{0,n}^0 > \frac{\psi}{2}\right] \\
&\leq \mathbb{P}\left[N_n^{(\infty)} > \frac{\psi}{2}\right] + \mathbb{P}\left[\sum_{i=1}^{N^0} 1[\theta_i^0 > \tau_n - \tau_0] > \frac{\psi}{2}\right] \\
&\leq \mathbb{P}\left[N_n^{(\infty)} > \frac{\psi}{2}\right] + \sum_{m=0}^{\infty} \mathbb{P}[N^0 = m]\mathbb{P}\left[\sum_{i=1}^{m} 1[\theta_i^0 > (1-\epsilon_1)n\mathbb{E}[\tau_1 - \tau_0]] > \frac{\psi}{2}\right] \\
&\qquad + \mathbb{P}[\tau_n - \tau_0 < (1-\epsilon_1)n\mathbb{E}[\tau_1 - \tau_0]], \qquad (24)
\end{aligned}$$

where in the previous inequalities $0 < \epsilon_1 < 1$ is an arbitrary constant and we used $N_n^{(\infty)} \geq N_{a,n}^{(C)}$ a.s., where $N_n^{(\infty)}$ is defined as in the proof of Theorem 1.

Now, we prove that there exists $\psi = \psi_0$ large enough such that (24) is bounded by $\epsilon/3$. By definition of $N_n^{(\infty)}$ in Section 2 and Little's formula, $\mathbb{E}N_n^{(\infty)} < \infty$ and, therefore,

$$\lim_{\psi\to\infty} \mathbb{P}\left[N^{(\infty)} > \frac{\psi}{2}\right] \to 0, \qquad (25)$$

uniformly for all $n > 0$. Next, note that $1[\theta_i^0 > (1-\epsilon_1)n\mathbb{E}[\tau_1 - \tau_0]] \leq 1[\theta_i^0 > (1-\epsilon_1)\mathbb{E}[\tau_1 - \tau_0]]$ a.s., and that for any fixed $m$,

$$\mathbb{P}\left[\sum_{i=1}^{m} 1[\theta_i^0 > (1-\epsilon_1)n\mathbb{E}[\tau_1 - \tau_0]] > \frac{\psi}{2}\right] \leq \mathbb{P}\left[\sum_{i=1}^{m} 1[\theta_i^0 > (1-\epsilon_1)\mathbb{E}[\tau_1 - \tau_0]] > \frac{\psi}{2}\right] \downarrow 0 \text{ as } \psi \to \infty,$$

which by the monotone convergence theorem implies that the second term in (24) satisfies

$$\lim_{\psi\to\infty} \sum_{m=0}^{\infty} \mathbb{P}[N^0 = m]\mathbb{P}\left[\sum_{i=1}^{m} 1[\theta_i^0 > (1-\epsilon_1)n\mathbb{E}[\tau_1 - \tau_0]] > \frac{\psi}{2}\right] = 0, \qquad (26)$$

uniformly for all $n > 0$. Finally, by the Weak Law of Large Numbers, for all $n$ large enough,

$$\mathbb{P}[\tau_n - \tau_0 \leq (1-\epsilon_1)n\mathbb{E}[\tau_1 - \tau_0]] \leq \epsilon/9. \qquad (27)$$

Thus, the previous conclusion in conjunction with (26), (25) and (24) implies that for an

arbitrary $0 < \epsilon < 1$, there exist $n_0 < \infty$ and $\psi_0 < \infty$ large enough such that for all $n \geq n_0$

$$\mathbb{P}\left[N^{(\infty)} > \frac{\psi_0}{2}\right] \leq \frac{\epsilon}{9} \text{ and } \sum_{m=0}^{\infty} \mathbb{P}[N^0 = m]\mathbb{P}\left[\sum_{i=1}^{m} 1[\theta_i^0 > (1-\epsilon_1)n\mathbb{E}[\tau_1 - \tau_0]] > \frac{\psi_0}{2}\right] \leq \frac{\epsilon}{9}. \tag{28}$$

Now, since $\mathbb{E}B_i < \infty$, $\mathbb{E}\theta_i < \infty$, there exist $\beta_0$, $\delta_0$, such that

$$\mathbb{P}[B_i > \beta_0] \leq \frac{\epsilon}{3\psi_0} \text{ and } \mathbb{P}[\theta_i > \delta_0] \leq \frac{\epsilon}{3\psi_0}.$$

Thus, the previous expressions in conjunction with (28), (27) and (23) imply that for all large enough $n \geq n_0$ inequality (22) holds for a chosen set $S(\psi_0, \beta_0, \delta_0)$.

Next, we show that there exists $n_1 > 0$ and $k > 0$ such that for all points $x \in S(\psi_0, \beta_0, \delta_0)$ and measurable sets $A \subset S(\psi_0, \beta_0, \delta_0)$, the following inequality holds

$$P_{n_1}(x, A) \geq kR(A). \tag{29}$$

Let $F_\theta(u)$ denote a cumulative distribution function of a random duration $\theta$, i.e., $\mathbb{P}[\theta \leq u]$. Furthermore, select a small positive number $\eta$ such that for some chosen $\Delta > \delta_0$, $F_\theta(\Delta) - F_\theta(\delta_0) = \eta > 0$. Next, for any $n_1$

$$P_{n_1}(x, A) \geq P_1(x, \omega_0)P_{n_2}(\omega_0, A), \tag{30}$$

where $n_2 = n_1 - 1$. Let $x = (m, b_1, \ldots, b_m, e_1, \ldots, e_m) \in S(\psi_0, \beta_0, \delta_0)$. Then,

$$\begin{aligned} P_1(x, \omega_0) &\geq \mathbb{P}[\tau_1 - \tau_0 \geq \Delta, \text{all } m \text{ requests depart in } (\tau_0, \tau_1)] \\ &= \int_{\Delta}^{\infty} \prod_{i=1}^{m} \frac{F_\theta(u + e_i) - F_\theta(e_i)}{1 - F_\theta(e_i)} dF_a(u), \end{aligned} \tag{31}$$

where $F_a(u)$ represents cumulative inter arrival distribution of a renewal process $\{\tau_n\}$, i.e., $F_a(u) = \mathbb{P}[\tau_1 - \tau_0 \leq u]$. Now, by applying lower bound

$$\frac{F_\theta(u + e_i) - F_\theta(e_i)}{1 - F_\theta(e_i)} \geq F_\theta(\Delta) - F_\theta(\delta_0) = \eta$$

in (31) we obtain

$$P_1(x,\omega_0) \geq \eta^m \mathbb{P}[\tau_1 - \tau_0 > \Delta] \geq \eta^{\psi_0}(1 - F_a(\Delta)). \tag{32}$$

Next, we derive a lower bound for $P_{n_2}(\omega_0, A)$ for some $n_2$ large enough such that

$$\mathbb{P}[\tau_{n_2} - \tau_0 > \delta] \geq 1 - \frac{\epsilon}{2}. \tag{33}$$

Note that the condition imposed on $n_2$ in (33) is possible due to the Weak Law of Large Numbers, since for any $\epsilon > 0$ and all $n_2$ large enough with $\delta_0 < (1-\epsilon)\mathbb{E}[\tau_{n_2} - \tau_0]$,

$$\mathbb{P}[\tau_{n_2} - \tau_0 > \delta_0] \geq \mathbb{P}[\tau_{n_2} - \tau_0 > (1-\epsilon)\mathbb{E}[\tau_{n_2} - \tau_0]] \geq 1 - \frac{\epsilon}{2},$$

Next, pick any $x' = (m', e'_1, \ldots, e'_{m'}, b'_1, \ldots, b'_{m'}) \in A$ where, without loss of generality, we assume that $e'_1 \geq e'_2 \geq \cdots \geq e'_{m'}$. Define $x' + dx' \triangleq (m', e'_1 + de'_1, \ldots, e'_{m'} + de'_{m'}, b'_1 + db'_1, \ldots, b'_{m'} + db'_{m'})$ where $de'_1, \cdots, de'_{m'}, db'_1, \cdots, db'_{m'}$ are infinitesimal elements. Then, the transition probability into state $(x', x' + dx')$ starting from $\omega_0$ can be bounded by the probability of the event that there are exactly $m'$ arrivals prior to $\tau_{n_2}$ whose arrival times are determined by $e'_1, \cdots, e'_{m'}$, whose resource requirements are in $(b'_1, b'_1 + db'_1), \cdots, (b'_{m'}, b'_{m'} + db'_{m'})$, and where the rest of $n_2 - m'$ arrivals are rejected since their requirements exceed capacity $C$. Therefore,

$$\begin{aligned}
P_{n_2}(\omega_0, (x', x' + dx')) &\geq \left\{ \prod_{j=1}^{m'} \mathbb{P}[\theta_j > e'_j] \mathbb{P}[B_j \in (b'_j, b'_j + db'_j)] \right\} \mathbb{P}[B_1 > C]^{n_2 - m'} \\
&\quad \times \mathbb{P}\left[ \bigcup_I \left\{ \sum_{j=i_1}^{i_2-1} Y_j \in (e'_1 - e'_2, e'_1 - e'_2 + de'_1), \cdots \sum_{j=i_{m'}}^{n_2-1} Y_j \in (e'_{m'}, e'_{m'} + de'_{m'}) \right\} \right] \\
&\geq \mathbb{P}[B_1 > C]^{n_2} \left\{ \prod_{j=1}^{m'} \mathbb{P}[\theta_j > e'_j] \mathbb{P}[B_j \in (b'_j, b'_j + db'_j)] \right\} \\
&\quad \times \mathbb{P}\left[ \bigcup_I \left\{ \sum_{j=i_1}^{i_2-1} Y_j \in (e'_1 - e'_2, e'_1 - e'_2 + de'_1), \ldots, \sum_{j=i_{m'}}^{n_2-1} Y_j \in (e'_{m'}, e'_{m'} + de'_{m'}) \right\} \right],
\end{aligned}$$

where $I \triangleq \{0 \le i_1 < i_2 < \cdots < i_{m'} \le n_2 - 1\}$ and $Y_j$ are i.i.d. random variables equal in distribution to inter-arrival times of the renewal process $\{\tau_n\}$, i.e., $Y_j \stackrel{d}{=} \tau_{j+1} - \tau_j$. Now denote

$$r(m', e'_1, \ldots, e'_{m'}, b'_1, \cdots, b'_{m'}) \triangleq \left\{ \prod_{j=1}^{m'} \mathbb{P}[\theta_j > e'_j] \mathbb{P}[B_j \in (b'_j, b'_j + db'_j)] \right\}$$
$$\times \mathbb{P}\left[ \bigcup_I \left\{ \sum_{j=i_1}^{i_2-1} Y_j \in (e'_1 - e'_2, e'_1 - e'_2 + de'_1), \ldots, \sum_{j=i_{m'}}^{n_2-1} Y_j \in (e'_{m'}, e'_{m'} + de'_{m'}) \right\} \right], \tag{34}$$

and define probability distribution

$$R(A) \triangleq \mathcal{V} \int_{x' \in A} r(m', e'_1, \ldots, e'_{m'}, b'_1, \ldots, b'_{m'}), \tag{35}$$

where $\mathcal{V}$ is a normalization constant. Note that $R(A)$ is well-defined since

$$\sum_{m'=0}^{\infty} \int_{\infty > e'_1 > \cdots > e'_{m'} > 0} \int_{b'_1, \ldots, b'_{m'} \ge 0}^{\infty} r(m', e'_1, \ldots, e'_{m'}, b'_1, \ldots b'_{m'})$$
$$\le \mathbb{P}[N_{0,n}^{(C)} \le \psi_0] (\mathbb{E}\theta_1)^{\psi_0} + \mathbb{P}[N_{0,n}^{(C)} > \psi_0]$$
$$\le \mathbb{P}[N_{0,n}^{(C)} \le \psi_0] (\mathbb{E}\theta_1)^{\psi_0} + \mathbb{P}[N_n^{(\infty)} > \psi_0] < \infty.$$

The previous inequalities, in conjunction with (34), (32) and (30) imply that

$$P_{n_2+1}(x, A) \ge \eta^{\psi_0} (1 - F_a(\Delta)) \mathbb{P}[B_1 > C]^{n_2} \mathcal{V}^{-1} R(A). \tag{36}$$

Finally, it is left to show that there exists $K > 0$ such that for every initial distribution $P_0$, for all $n$ large and for any measurable set $A \subset S(\psi_0, \beta_0, \delta_0)$

$$P_n(A) \le K R(A) + \epsilon.$$

By (33), for all $n \geq n_2$

$$\begin{aligned}
P_n(A) &\leq \mathbb{P}[X_n \in A, \tau_n - \tau_0 > \delta_0] + \mathbb{P}[\tau_n - \tau_0 \leq \delta_0] \\
&\leq \mathbb{P}[X_n \in A, \tau_n - \tau_0 > \delta_0] + \epsilon/2 \\
&\leq \int_{x' \in A} \left\{ \prod_{j=1}^{m'} \mathbb{P}[\theta_j > e'_j] \mathbb{P}[B_j \in (b'_j, b'_j + db'_j)] \right\} \\
&\qquad \times \mathbb{P}\left[ \bigcup_I \left\{ \sum_{j=i_1}^{i_2-1} Y_j \in (e'_1 - e'_2, e'_1 - e'_2 + de'_1), \ldots, \sum_{j=i_{m'}}^{n_2-1} Y_j \in (e'_{m'}, e'_{m'} + de'_{m'}) \right\} \right] + \epsilon \\
&= \int_{x' \in A} r(m', e'_1, \ldots, e'_{m'}, b'_1, \ldots, b'_{m'}) + \epsilon,
\end{aligned}$$

where the second inequality follows from the fact that requests that are active at $\tau_n$ must occur in the previous $\delta_0$ length of time that are captured in $n_2$ renewal intervals $[\tau_{n-n_2}, \tau_{n-n_2+1}), \ldots, [\tau_{n-1}, \tau_n)$ with significant probability (greater than $1 - \epsilon/2$). Thus, after applying definition (35), we obtain that for all $n$ large

$$P_n(A) \leq \mathcal{V}^{-1} R(A) + \epsilon,$$

which, in conjunction with (36) and (22), implies that the process $X_n$ satisfies conditions of the theorem stated at the beginning of this section. Thus, there exists a unique stationary distribution for the Markov chain $X_n$. Therefore, since $Q_n^{(C)}$ defined in Section 2 is a functional of the process $X_n$, it has a unique stationary distribution as well implying the existence of the stationary blocking probability.

$\diamond$

**Acknowledgments**

The authors would like to thank Prof. Predrag Jelenković for valuable suggestions related to the possible generalizations of this work.

**References**

[1] ASMUSSEN, S., HENRIKSEN, L. F. AND KLÜPPELBERG, C. (1994). Large claims approximations for risk processes in a Markovian environment. *Stochastic Processes and their Applications* **54,** 29–43.

[2] ASMUSSEN, S. AND PIHLGÅRD, M. Loss rates for Lévy processes with two reflecting barriers. *Mathematics of OR*. to appear.

[3] ATHREYA, K. B. AND NEY, P. E. (1972). *Branching Processes*. Springer-Verlag.

[4] BONALD, T. (2006). The Erlang model with non-Poisson call arrivals. *Procedings of ACM/Sigmetric and Performance Conference*, 276–286.

[5] EMBRECHTS, P., K. C. AND MIKOSCH, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer.

[6] EMBRECHTS, P. AND GOLDIE, C. M. (1980). On closure and factorization properties of subexponential and related distributions. *J. Austral. Math. Soc.* **Series(A) 29,** 243–256.

[7] ERLANG, A. K. (1917). Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Elektrotkeknikeren* **13,** 5–13.

[8] GAZDZICKI, P., LAMBADARIS, I. AND MAZUMDAR, R. R. (1993). Blocking probabilities for large multirate Erlang loss systems. *Advances in Applied Probability* **25,** 997–1009.

[9] GOLDIE, C. M. AND KLÜPPELBERG, C. (1998). Subexponential distributions. In *A Practical Guide to Heavy Tails: Statistical Techniques for Analysing Heavy Tailed Distributions*. ed. M. T. R. Adler, R. Feldman. Birkhäuser, Boston pp. 435–459.

[10] HEYMAN, D. P. AND LAKSHMAN, T. V. (1996). Source models for VBR broadcast-video traffic. *IEEE/ACM Transactions on Networking* **4,** 40–48.

[11] JELENKOVIĆ, P., LAZAR, A. AND SEMRET, N. (1995). Multiple time scales and subexponentiality in MPEG video streams. *Technical report* CU/CTR/TR 430-95-36. Columbia University. http://www.ctr.columbia.edu/comet/publications.

[12] JELENKOVIĆ, P. R. (1999). Subexponential loss rates in a GI/GI/1 queue with applications. *Queueing Systems, Special Issue on Long-Tailed Distributions* **33,** 91–123.

[13] KAUFMAN, J. S. (1981). Blocking in a shared resources environment. *IEEE Transactions on Communications* **29,** 1474–1481.

[14] KELLY, F. P. (1986). Blocking probabilities in large circuit-switched networks. *Advances in Applied Probability* **18,** 473–505.

[15] KELLY, F. P. (1991). Loss networks. *Annals of Applied Probability* **1,** 319–378.

[16] KRISHNAN, K. R. AND MEEMPAT, G. (1997). Long-range dependence in VBR video streams and atm traffic engineering. *Performance Evaluation* **30,** 46–56.

[17] LIU, L., KASHYAP, B. R. K. AND TEMPLETON, J. G. C. (1990). On the $GI^X/G/\infty$ system. *Journal of Applied Probability* **27,** 671–683.

[18] LOUTH, G., MITZENMACHER, M. AND KELLY, F. (1994). Computational complexity of loss networks. *Theoretical Computer Science* **125,** 45–59.

[19] LU, Y., RADOVANOVIĆ, A. AND SQUILLANTE, M. (2006). Workforce management through stochstic network models. *Proceedings of IEEE SOLI Conference*.

[20] SEVASTYANOV, B. A. (1957). An ergodic theorem for Markov processes and its application to telephone systems with refusals. *Theory of probability and its applications* **2,** 104–112.

[21] TAKACS, L. (1980). Queues with infinitely many servers. *R.A.I.R.O. Recherche Operationnelle* **14,** 109–113.

[22] WHITT, W. (1985). Blocking when service is required from several facilities simultaniously. *AT&T Technical Journal* **64,** 1807–1856.

[23] WILKINSON, R. I. (1956). Theory of toll traffic engineering in the USA. *Bell System Technical Journal* **35,** 421–513.

[24] WOLFF, R. W. (1989). *Stochastic Modeling and Theory of Queues*. Prentice Hall.

[25] ZACHARY, S. (1991). On blocking in loss networks. *Advances in Applied Probability* **23,** 355–372.